Epidemics 17 (2016) 1–9

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Research Paper

# Clustering of contacts relevant to the spread of infectious disease

Xiong Xiao[a,d], Albert Jan van Hoek[a], Michael G. Kenward[a], Alessia Melegaro[b], Mark Jit[a,c,*]

[a] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom
[b] DONDENA Centre for Research on Social Dynamics & Public Policy, Università Bocconi, Via Guglielmo Röntgen n. 1, 20136 Milan, Italy
[c] Modelling and Economics Unit, Public Health England, 61 Colindale Avenue, London NW9 5EQ, United Kingdom
[d] Department of Epidemiology and Biostatistics, West China School of Public Health, Sichuan University, Chengdu, China

ARTICLE INFO

*Article history:*
Received 18 January 2016
Received in revised form 4 August 2016
Accepted 23 August 2016
Available online 26 August 2016

*Keywords:*
Clustering
Contacts
Infectious diseases
Mathematical modelling
Varicella-zoster virus

ABSTRACT

*Objective:* Infectious disease spread depends on contact rates between infectious and susceptible individuals. Transmission models are commonly informed using empirically collected contact data, but the relevance of different contact types to transmission is still not well understood. Some studies select contacts based on a single characteristic such as proximity (physical/non-physical), location, duration or frequency. This study aimed to explore whether clusters of contacts similar to each other across multiple characteristics could better explain disease transmission.
*Methods:* Individual contact data from the POLYMOD survey in Poland, Great Britain, Belgium, Finland and Italy were grouped into clusters by the *k* medoids clustering algorithm with a Manhattan distance metric to stratify contacts using all four characteristics. Contact clusters were then used to fit a transmission model to sero-epidemiological data for varicella-zoster virus (VZV) in each country.
*Results and discussion:* Across the five countries, 9–15 clusters were found to optimise both quality of clustering (measured using average silhouette width) and quality of fit (measured using several information criteria). Of these, 2–3 clusters were most relevant to VZV transmission, characterised by (i) 1–2 clusters of age-assortative contacts in schools, (ii) a cluster of less age-assortative contacts in non-school settings. Quality of fit was similar to using contacts stratified by a single characteristic, providing validation that single stratifications are appropriate. However, using clustering to stratify contacts using multiple characteristics provided insight into the structures underlying infection transmission, particularly the role of age-assortative contacts, involving school age children, for VZV transmission between households.


## 1. Introduction

Mathematical models of infectious disease transmission require assumptions about mixing between different subgroups in a population that can potentially lead to transmission between infected and susceptible individuals. The simplest assumption is that everyone has the same probability of contacting each other, but this can sometimes lead to misleading results (Keeling and Rohani, 2008). Indeed, many infection control interventions such as vaccinating children (Thorrington et al., 2015) or closing schools during a pandemic (House et al., 2011) are predicated on the assumption that certain subgroups in the population are the main transmitters.

A more realistic assumption is to subdivide the population based on some characteristic, and introduce a matrix of contact rates capable of transmitting infection between each subgroup, called the "who acquires infection from whom" (WAIFW) matrix (Vynnycky and White, 2010). Age is the characteristic most commonly used as a source of heterogeneity in mixing patterns. A model with age-stratified contact rates can be fitted to age-specific data on infection history (such as sero-epidemiological data, which marks the prevalence of previous infection) to estimate the age-specific effective contact rates in the WAIFW matrix.

To inform the elements of the WAIFW matrix, the number of social contacts that individuals in different age groups report can

*Abbreviations:* AIC, Akaike Information Criterion; AICc, small-sample-size corrected Akaike Information Criterion; ASW, average silhouette width; BIC, Bayesian Information Criterion; PAM, partitioning around medoids; VZV, varicella-zoster virus; WAIFW, who acquires infection from whom.

* Corresponding author at: Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom.
*E-mail addresses:* Xiong.Xiao@lshtm.ac.uk (X. Xiao), Albert.VanHoek@lshtm.ac.uk (A.J. van Hoek), Mike.Kenward@lshtm.ac.uk (M.G. Kenward), alessia.melegaro@unibocconi.it (A. Melegaro), mark.jit@lshtm.ac.uk (M. Jit).

http://dx.doi.org/10.1016/j.epidem.2016.08.001

be empirically measured and used as proxies to the actual contact rates underlying transmission (Beutels et al., 2006; Edmunds et al., 1997; Wallinga et al., 2006). The largest such study is a diary-based survey of 7290 participants in eight European countries collected in 2006 as part of the POLYMOD project (Mossong et al., 2008). Since then contact studies have been carried out in other parts of the world using similar methodology.

In these studies, participants are asked to record the contacts that they have made over a single day and classify them using a number of characteristics (such as physical/non-physical, long/short, home/school/work etc.) Since it is unrealistic to measure effective contacts (i.e. contacts that can transmit a particular infection) between individuals directly, some form of self-reported social behaviour such as face-to-face conversation or skin-to-skin contact is used as a proxy. It is assumed that the age distribution of these social contacts is related to the age distribution of effective contacts by a constant proportionality factor, an assumption referred to as the "social contact hypothesis" (Wallinga et al., 2006). Hence the age-specific transmission matrix is completely described by estimating this factor. Subsequent analyses (Melegaro et al., 2011) found that stratifying the WAIFW matrix according to different characteristics of contacts (with a different proportionality constant for each type of contact) significantly improved the goodness of fit of the models to serological markers of past infection for respiratory infections. This implies that different characteristics of social contact contribute differently to infection transmission.

A previous study explored the use of formal clustering algorithms to group POLYMOD survey respondents based on the number and location of their contacts (Kretzschmar and Mikolajczyk, 2009). The study found that respondents across different countries fell into a similar range of contact profiles, with the work, school and household contact profiles most common. However, this does not tell us whether there are certain types of contacts (rather than respondents) which may be particularly relevant for the transmission of particular infectious diseases. Furthermore, the relevant clusters of contacts may be disease-specific, since different infections have separate routes of transmission.

Hence an alternative approach would be to explore what kind of contacts (rather than respondents) are most relevant to infection transmission. The only previous work in this area has focused on single dimensions of contacts (Melegaro et al., 2011). This simply indicated that "intimate" (i.e. physical, home, long-duration and frequent) contacts are better able to explain age-dependent patterns in the acquisition of serological markers for varicella-zoster virus (VZV) and parvovirus B19, the two infection examined in the study. Since the dimensions of intimacy are highly correlated, it may be more informative to take all the characteristics of social contacts into account collectively when stratifying the WAIFW matrix.

In particular contacts made by different respondents could be grouped into clusters, and then examined to identify which clusters best explain patterns of infection acquisition in the corresponding populations. Here, we explore the use of clustering algorithms to determine clusters of social contacts which are similar to each other based on multi-dimensional characteristics of social contacts.

A range of clustering algorithms have been developed such as hierarchical clustering, partitioning clustering and latent class clustering (Everitt et al., 2011). In hierarchical clustering, each element is plotted on a graph to determine the optimal number of clusters. When the number of elements becomes very large (such as the thousands of contacts for each country in the POLYMOD survey), this graphical method becomes very cumbersome. In latent class clustering, a multivariate distribution is imposed on the data and the validation of the clustering results depends on several assumptions such as the parametric form of the multivariate distribution, the dependency between variables and the approximation of likelihood estimation. Since social contact data include different types of variables (binary and ordinal), some of which are structured, it is difficult to identify an appropriate multivariate distribution to describe the data. Hence we used partitioning clustering, and in particular the *k* medoids method to classify contacts.

We then investigate whether these clusters enhance our understanding of transmission patterns by using them to fit a transmission dynamic model to age-dependent patterns in the acquisition of varicella-zoster virus serological markers, as an example of a childhood respiratory infection with clear, long-lived markers of past infection and no vaccination history at the time of data collection.

## 2. Methods

### 2.1. Data sources

Age-specific contact matrices were constructed using social contact data from participants of the POLYMOD project (Mossong et al., 2008) living in Poland (15808 contacts, 1003 participants), Great Britain (11052 contacts, 996 participants), Belgium (8810 contacts, 747 participants), Finland (10319 contacts, 973 participants) and Italy (15788 contacts, 842 participants). Contacts with any missing information were discarded, so our dataset differs slightly from previous analyses (Melegaro et al., 2011). Contact matrices were adjusted for population size and reciprocity using well-described procedures (Melegaro et al., 2011; Wallinga et al., 2006).

Models were fitted to data on the presence of antibodies to VZV from serum samples collected in 1996 from 1300 participants aged 0–19 in Poland, 2091 participants aged 0–20 in England and Wales (fitted to Great British contacts), 2760 participants aged 0–39 in Belgium, 2500 participants aged 0–79 in Finland and 2517 participants aged 0–79 in Italy (Vyse et al., 2004). The samples were collected from unlinked anonymised (apart from age) residual sera following microbiological or biochemical investigations (Osborne et al., 2000), and tested as part of the European Commission-funded second European Sero-epidemiological Network (ESEN2) (Melegaro et al., 2011; Nardone et al., 2007). Children under 5 years old were oversampled with the sample size in each age group ranging from 117 to 192. For those aged 5–20 years approximately 100 sera in each one-year age group were tested. All serological tests for VZV-specific IgG were performed at Preston Public Health Laboratory using a commercial ELISA assay.

Population data were obtained from national statistics offices in the five countries considered as in previous analyses (Melegaro et al., 2011).

### 2.2. Cluster analysis of social contact data

A cluster is defined as a group of social contacts whose members are more similar to each other than to non-members. The similarity of two contacts is defined using the Manhattan distance measure between the two (Everitt et al., 2011), based on four contact characteristics: proximity (physical, non-physical), duration (<5 min, 5–15 min, 15 min to 1 h, 1–4 h, >4 h), frequency (daily, weekly, monthly, a few times a year, first time) and location (home, work, school, leisure, transport, other). Variables representing each characteristic were recoded so that the distance between any two contacts lies between 0 and 1, so the Manhattan distance becomes the Gower's similarity coefficient which is an appropriate proximity measurement for mixed data (Gower, 1971) (see Appendix A1 for details in Supplementary data). Contacts recorded as taking place in multiple locations were assigned to a single location based on the following hierarchy: home > work > school > leisure > other > transport. The

hierarchy was based on the putative duration of contacts in the given settings as suggested by previous investigators (Kretzschmar and Mikolajczyk, 2009). Social contacts were classified into clusters using the Partitioning Around Medoids (PAM) algorithm, the most common realisation of *k*-medoids clustering. This approach pre-defines a fixed number of clusters, selects typical elements as centroids of each cluster and assigns other elements to a cluster according to their distance to the centroid (Kaufman and Rousseeuw, 1990).

The number of clusters was varied between 2 and 15, with the upper limit included to avoid possible problems with data sparsity. The optimal number of clusters for the PAM algorithm was then determined using average silhouette width (*ASW*), which measures how well each contact is assigned to its cluster (Kaufman and Rousseeuw, 1990). Replication analysis (Breckenridge, 2000; Walesiak et al., 2008) was performed to assess the robustness of the clustering results, by splitting the dataset into several subsets and using the same algorithm separately for each subset to compare classification agreement.

### 2.3. Fitting dynamic transmission models

We constructed a realistic age-structured (Schenzle, 1984) susceptible-infected-recovered (SIR) dynamic model of VZV transmission (with no natural mortality assumed). In this model, the next generation matrix representing the potential number of infection transmission events per person, is calculated as the product of the (adjusted) age-dependent contact matrix, a proportionality factor $q$ representing the proportion of contacts that can potentially lead to a new infection and vector representing the size of the susceptible population in each age group. For each cluster $i$ of contacts ($1 \leq i \leq$ N where N is the total number of clusters), we generated $N$ separate contact matrices $C_i$ with corresponding proportionality factors $q_i$. Hence the next generation matrix is the linear combination of all contact matrices $C_1\ q_1\ w + \ldots + C_N\ q_N\ w$ where $w$ is a vertical vector of the population size in each age group. The basic reproduction $R_0$ was calculated as the principal eigenvalue of the next generation matrix.

A grid search algorithm was used to find the $q_i$'s which maximised the binomial likelihood of the model given sero-epidemiological data for VZV infections (Melegaro et al., 2011; Ogunjimi et al., 2009). We defined relevant clusters as those with corresponding best fitting $q_i$'s that were non-zero, and hence which contributed to the next generation matrix. An iterative method was used to obtain the proportion of immunes at each age group as a function of the $q_i$'s. Quality of model fit with different numbers of clusters was measured using the Akaike Information Criterion (*AIC*), its small-sample-size corrected version (*AICc*) and the Bayesian Information Criterion (*BIC*). Goodness of fit was also compared to models with contacts stratified by a single characteristic only.

### 2.4. Uncertainty intervals

Uncertainty intervals for the proportionality factor $q$ and basic reproduction number $R_0$ were estimated by bootstrap sampling from both social contact data and sero-epidemiological data with the same sample size as the original data (Melegaro et al., 2011). For social contact data, bootstrap samples were generated by re-sampling participants' identity numbers. Social contact matrices were then estimated for each bootstrap sample. For the sero-epidemiological data, bootstrap samples were generated by randomly drawing from a Bernoulli distribution with probability of success equal to the observed seroprevalence for the specific age group. After this, the newly generated social contact matrices and sero-epidemiological data were matched to repeatedly estimate the transmission parameters.

All analyses were performed in R version 2.15.2 (R Core Team, 2012), using the *cluster* and *clusterSim* packages.

### 2.5. Sensitivity analysis

To explore whether between-country differences in fitted models were due to heterogeneity in the age ranges over which seroprevalence data was available, we repeated our analysis using only seroprevalence data in the range 0–20 years (which is the maximum range for which data were available in all five countries).

## 3. Results

### 3.1. Optimal number of clusters

Evaluating the optimal number of clusters involves three components: the quality of the clustering (measured using *ASW*), how well the resulting clusters fit the transmission model of VZV seroprevalence (measured using *AIC*, *AICc* and *BIC*) and the number of clusters that are actually relevant to this model fit (Fig. 1).

In all countries, ASW decreased between 2 and 5 clusters (indicating worse clustering quality), probably because most contact characteristics have either 2 (physical/non-physical) or 5 (other characteristics) levels of encoding. After 5 clusters, the ASW gradually increased, with the increases occurring at the points where the number of relevant clusters increased. The three measures of goodness of fit to VZV seroprevalence data (*AIC*, *AICc* and *BIC*) were highly consistent with each other.

We determined the optimal number of clusters by maximizing both quality of the clustering as well as goodness of fit to the point where further increases in the number of clusters would not make noticeable changes in the quality of clustering, model fits or the number of relevant clusters substantially. The optimal number of clusters is 12 in Great Britain, 14 in Italy, 15 in Belgium, 8 in Finland and 7 in Poland. The related corrected Rand index for the clustering in each country ranges from 89% to 95%, which suggests that the clustering is robust (Breckenridge, 2000). The Rand index represents the average agreement between two data clustering in multiple replications (with ten replications used in this analysis) if contact data are randomly divided into two samples and each sample is clustered into same number of clusters separately.

### 3.2. Comparisons between different stratifications

When contact data are stratified by clusters, only 2–3 clusters in each country are found to be relevant to VZV transmission (i.e. have non-zero $q_i$'s when the model is fitted to VZV serology). Previous work stratifying contacts by single characteristic suggested that more intimate contacts (physical, home/school, frequent, long duration) best explain VZV transmission (Melegaro et al., 2011). When clustering contacts using multiple characteristics, the most relevant clusters are still likely to contain more intimate contacts. However, the picture is more nuanced and varies across the five countries.

Details of the most relevant clusters are given in Table 2, Fig. 2, Fig. 3 and Appendix A.2 (Supplementary data). Briefly, in Finland and Italy, the countries with the most strongly age-assortative overall contact matrices, three contact clusters are relevant. Overall assortativity is driven by two clusters dominated by daily school contacts: one physical and one non-physical. A third cluster, dominated by non-physical leisure contacts in Finland and physical other contacts in Italy, is also assortative but less so than the school contacts. In both countries there are two clusters with home contacts which are not found to be relevant.

In Great Britain, Belgium and Poland, only two clusters are relevant: one consisting mainly of highly assortative daily physical

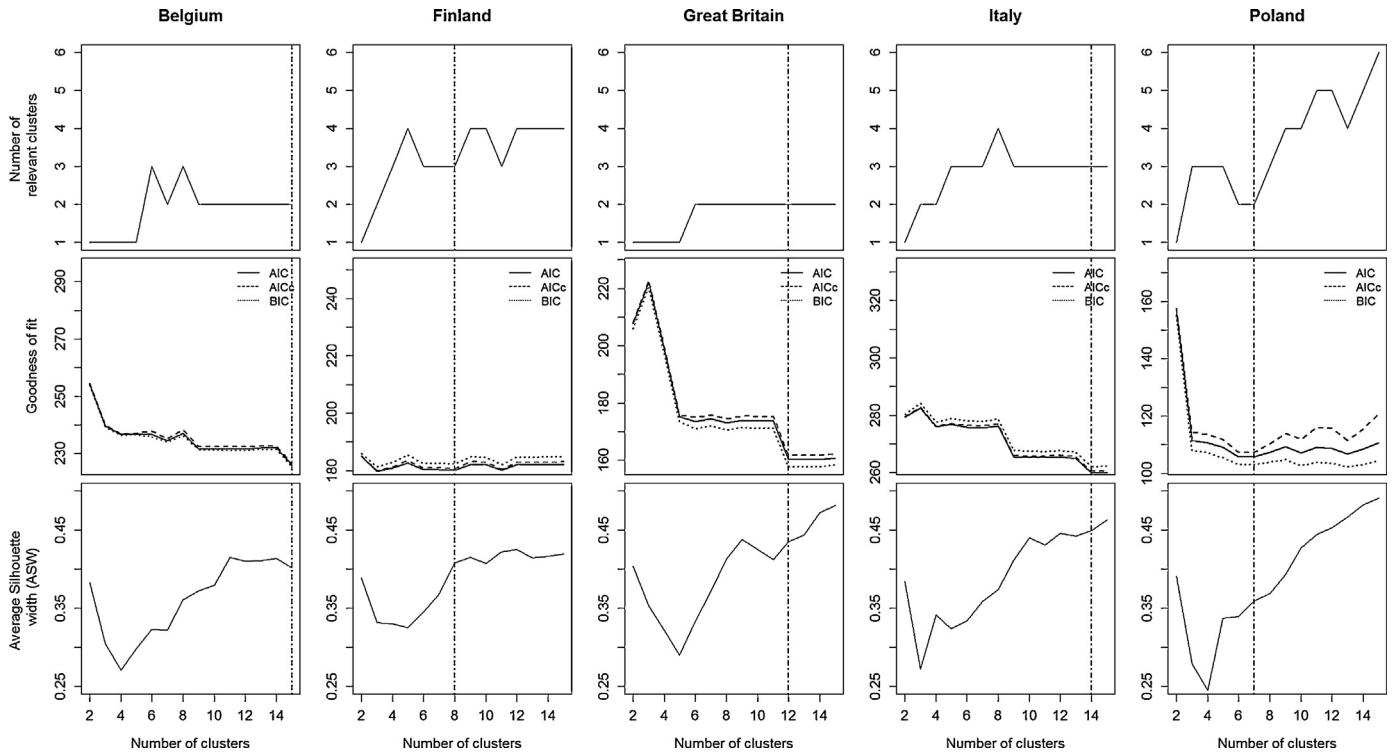


**Fig. 1.** Quality of clustering (measured using ASW), quality of fit (measured using AIC, AICc and BIC) and number of relevant clusters for models in each country with different number of clusters. Vertical dashed line shows the model with the optimal number of clusters.

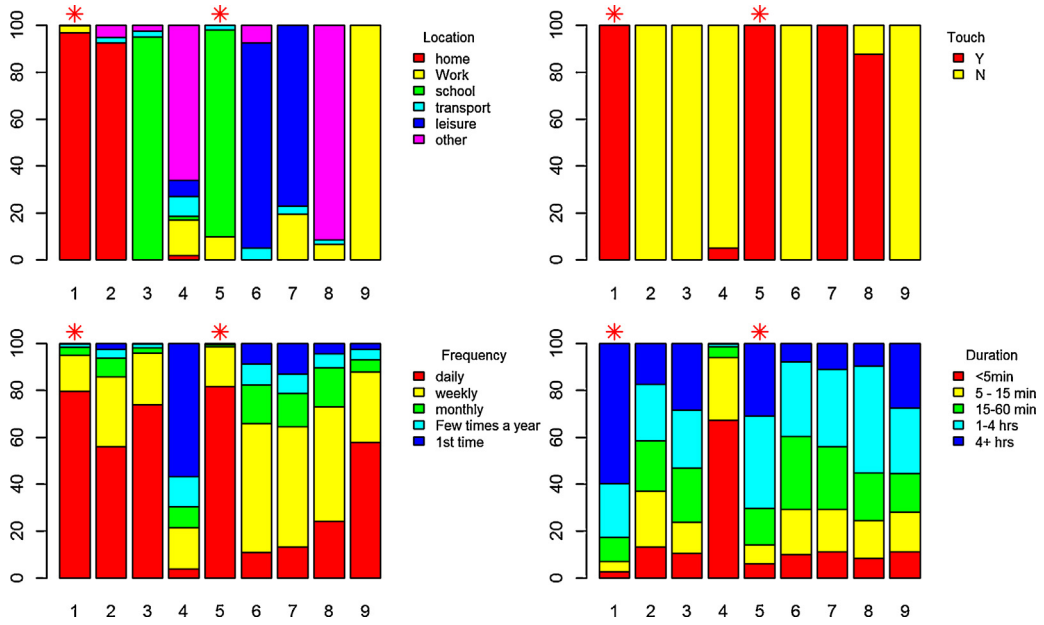


**Fig. 2.** The distributions of the characteristics of contact for each cluster in Great Britain. Each cluster corresponds to one numbers below the x-axis. The red star mark indicates the relevant cluster whose estimated $q$ is not equal to zero. Results for other countries are given in Appendix A2 (Supplementary data). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

school contacts, and the second of less assortative physical home contacts. The home clusters show a strong secondary diagonal probably representing inter-generational contacts, particularly in Poland. In all three countries, there are non-relevant home and school clusters more dominated by longer (>4 h) contacts than the home and school clusters actually found relevant.

Table 1 compares the best fitting model with the optimal number of clusters in each country, with models of contacts stratified by single characteristics as used in previous analyses (Melegaro et al., 2011). In every country, a model based on clustering contacts using multiple characteristics has a better quality fit (measuring using AIC) than models with contacts stratified by a single characteristics. However, the improved fit is at the expense of poorer model

**Table 1**
Model outcomes for different stratifications of social contact data. Parantheses give 95% interval for the bootstrap distributions of $q$ and $R_0$.

| | Great Britain | | Italy | | Belgium | | Finland | | Poland | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 95% CI | | 95% CI | | 95% CI | | 95% CI | | 95% CI |
| All contacts | | | | | | | | | | |
| q | 0.057 | 0.048–0.066 | 0.033 | 0.027–0.039 | 0.091 | 0.068–0.114 | 0.076 | 0.065–0.089 | 0.057 | 0.046–0.071 |
| R0 | 4.65 | 3.95–5.50 | 4.59 | 3.74–5.64 | 7.72 | 5.81–10.31 | 5.84 | 4.95–7.01 | 6.87 | 5.49–8.90 |
| AIC | 264 | | 295 | | 277 | | 194 | | 181 | |
| By proximity | | | | | | | | | | |
| q (physical) | 0.090 | 0.073–0.107 | 0.043 | 0.026–0.052 | 0.114 | 0.083–0.142 | 0.030 | 0.000–0.095 | 0.078 | 0.058–0.095 |
| q (non-physical) | 0.000 | 0.000–0.011 | 0.002 | 0.000–0.036 | 0.000 | 0.000–0.036 | 0.154 | 0.056–0.221 | 0.000 | 0.000–0.055 |
| R0 | 3.70 | 3.29–4.40 | 3.55 | 3.12–4.71 | 5.74 | 4.46–7.64 | 8.34 | 5.78–11.19 | 5.47 | 4.47–7.43 |
| AIC | 202 | | 273 | | 253 | | 182 | | 158 | |
| By duration | | | | | | | | | | |
| q (<15 min) | 0.000 | 0.000–0.000 | 0.000 | 0.000–0.020 | 0.000 | 0.000–0.001 | 0.000 | 0.000–0.287 | 0.000 | 0.000–0.121 |
| q (15–60 min) | 0.000 | 0.000–0.000 | 0.000 | 0.000–0.130 | 0.000 | 0.000–0.535 | 0.000 | 0.000–0.117 | 0.000 | 0.000–0.000 |
| q (>1 h) | 0.081 | 0.060–0.096 | 0.043 | 0.018–0.050 | 0.121 | 0.026–0.148 | 0.107 | 0.042–0.131 | 0.079 | 0.027–0.103 |
| R0 | 3.87 | 3.28–5.63 | 3.89 | 3.37–5.20 | 5.39 | 4.15–9.20 | 4.51 | 3.97–9.01 | 5.27 | 4.31–8.15 |
| AIC | 235 | | 285 | | 250 | | 186 | | 186 | |
| By location | | | | | | | | | | |
| q (home) | 0.185 | 0.092–0.213 | 0.021 | 0.000–0.074 | 0.105 | 0.000–0.305 | 0.044 | 0.000–0.131 | 0.224 | 0.082–0.326 |
| q (school) | 0.000 | 0.000–0.111 | 0.044 | 0.023–0.060 | 0.154 | 0.000–0.298 | 0.142 | 0.068–0.191 | 0.025 | 0.000–0.049 |
| q (work) | 0.000 | 0.000–0.000 | 0.000 | 0.000–0.012 | 0.000 | 0.000–0.000 | 0.000 | 0.000–0.000 | 0.000 | 0.000–0.000 |
| q (others) | 0.000 | 0.000–0.337 | 0.008 | 0.000–0.065 | 0.000 | 0.000–0.120 | 0.026 | 0.000–0.179 | 0.000 | 0.000–0.229 |
| R0 | 5.11 | 3.93–8.93 | 3.93 | 3.36–5.07 | 4.97 | 3.98–8.54 | 6.15 | 4.35–8.56 | 7.63 | 6.07–10.77 |
| AIC | 192 | | 281 | | 251 | | 183 | | 108 | |
| By frequency | | | | | | | | | | |
| q (daily) | 0.029 | 0.000–0.056 | 0.039 | 0.025–0.046 | 0.162 | 0.076–0.220 | 0.120 | 0.051–0.144 | 0.089 | 0.062–0.115 |
| q (weekly) | 0.155 | 0.057–0.268 | 0.000 | 0.000–0.094 | 0.000 | 0.000–0.129 | 0.015 | 0.000–0.405 | 0.000 | 0.000–0.279 |
| q (less than weekly) | 0.000 | 0.000–0.000 | 0.000 | 0.000–0.085 | 0.000 | 0.000–0.101 | 0.000 | 0.000–0.082 | 0.000 | 0.000–0.000 |
| R0 | 4.62 | 3.68–6.60 | 4.19 | 3.50–5.33 | 4.83 | 3.80–7.50 | 5.35 | 4.60–7.17 | 5.53 | 4.03–9.82 |
| AIC | 166 | | 282 | | 231 | | 183 | | 187 | |
| By clusters | | | | | | | | | | |
| Clusters | 12 | | 14 | | 15 | | 8 | | 7 | |
| Relevant clusters | 2 | | 3 | | 2 | | 3 | | 2 | |
| q (first cluster) | 0.565 | 0.000–0.723 | 0.071 | 0.000–0.225 | 0.204 | 0.000–0.336 | 0.303 | 0.000–0.371 | 0.246 | 0.000–0.323 |
| q (second cluster) | 0.085 | 0.000–0.178 | 0.127 | 0.000–0.202 | 0.443 | 0.000–0.948 | 0.183 | 0.000–0.329 | 0.081 | 0.000–0.129 |
| q (third cluster) | | | 0.064 | 0.000–0.105 | | | 0.030 | 0.000–0.357 | | |
| R0 | 3.79 | 2.79–9.44 | 4.68 | 2.92–5.36 | 4.49 | 3.22–8.25 | 6.19 | 4.92–23.29 | 7.61 | 5.91–32.33 |
| AIC | 160 | | 260 | | 226 | | 180 | | 106 | |
| Rand index | 0.89 | | 0.91 | | 0.91 | | 0.95 | | 0.90 | |

**Table 2**
Characteristics of the most relevant clusters in each country.

| Country | Cluster | Location | Physical | Frequency | Duration | Assortativity | Importance |
|---|---|---|---|---|---|---|---|
| Great Britain | 1 (home) | Home | Y | Mixed | Mixed | – | ++ |
| | 2 (school) | School | Y | Daily | Mixed | + | + |
| Belgium | 1 (home) | Home | Y | Daily | Mixed | – | – |
| | 2 (school) | School | Y | Daily | Mixed | + | + |
| Poland | 1 (home) | Home | Y | Daily | Mixed | – | ++ |
| | 2 (school) | School (mostly) | Y | Daily (mostly) | Mixed | + | + |
| Finland | 1 (school touch) | School | Y | Daily (mostly) | Mixed | ++ | + |
| | 2 (school non-touch) | School | N | Daily (mostly) | Mixed | ++ | – |
| | 3 (leisure) | Leisure | N | Mixed | Mixed | + | + |
| Italy | 1 (school touch) | School | Y | Daily (mostly) | Mixed | ++ | + |
| | 2 (school non-touch) | School | N | Daily (mostly) | Mixed | ++ | + |
| | 3 (other) | Other | Y | Mixed | Mixed | + | – |

estimations (larger uncertainty intervals) around the estimated $R_0$. The models based on clusters with multiple characteristics estimate that $R_0$ ranges from 3.79 (95% range 2.79–9.44) in Great Britain to 7.61 (95% range 5.91–32.33) in Poland. The point estimates are similar to estimates using single characteristics and consistent with a range of 3–8 in the literature (Goeyvaerts et al., 2010; Iozzi et al., 2010; Melegaro et al., 2011; Ogunjimi et al., 2009).

Fig. 4 shows model predictions of VZV seropositivity using the optimal number of clusters compared them with the true proportion of seropositive samples. On the whole model predictions fit the sero-epidemiological data well.

### 3.3. Sensitivity analysis

When only seroprevalence data for 0–20 year olds is used in Belgium, Finland and Italy to ensure comparability with Great Britain and Poland, the overall results are not greatly altered. The same three highly assortative clusters (dominated by two clusters

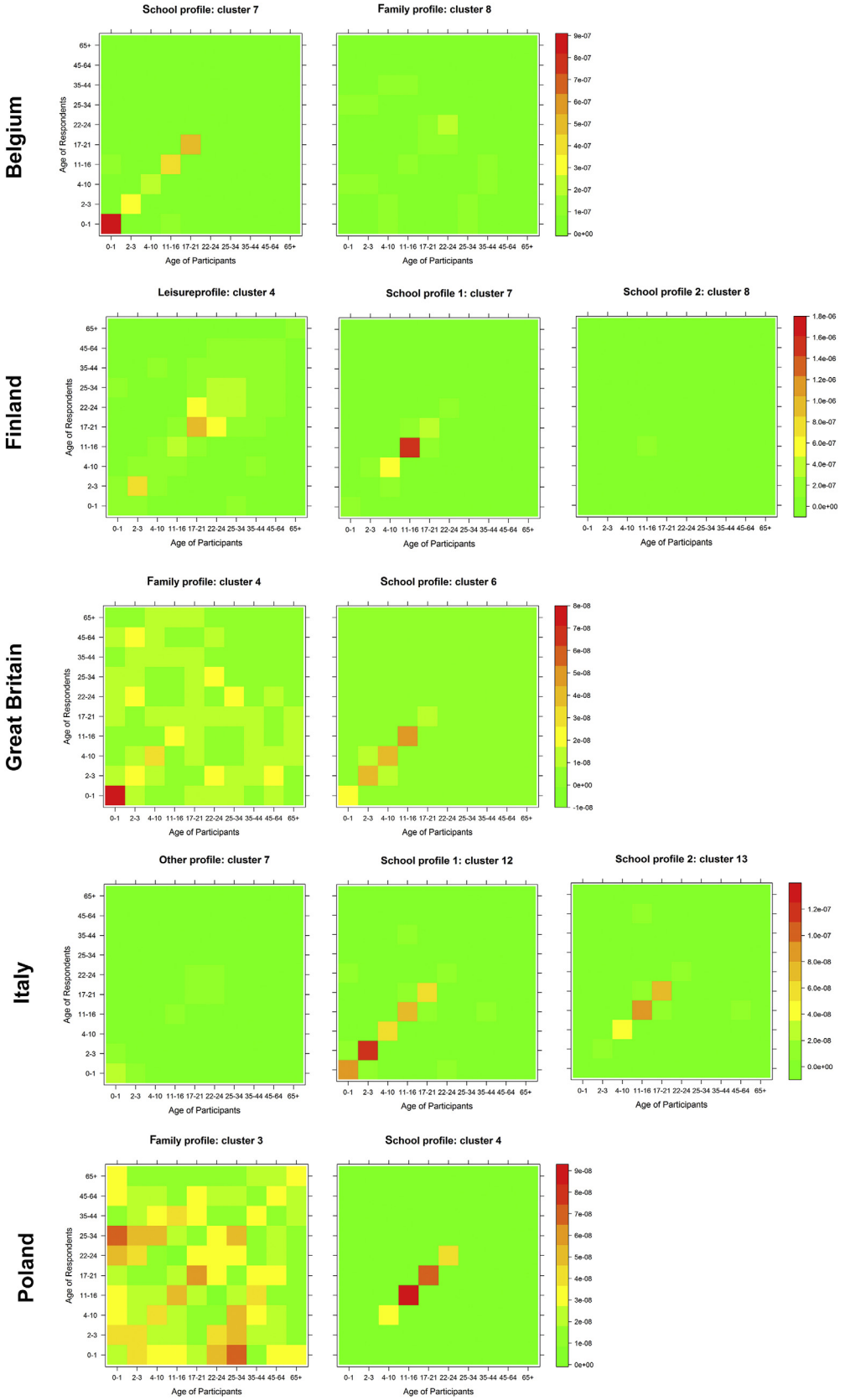


**Fig. 3.** The corrected age-specific contact matrices for the two identified relevant clusters (the cluster number 1 and 5) with sampling weighting and reciprocal correction.

of school contacts) are relevant in Italy. In Finland, two assortative clusters (dominated by school and leisure contacts respectively) are relevant, so there is little overall change in age-assortativity. In Belgium, there is a slight shift as the highly assortative cluster dominated by school contacts is dropped but a less assortative cluster dominated by home contacts remains relevant, so overall

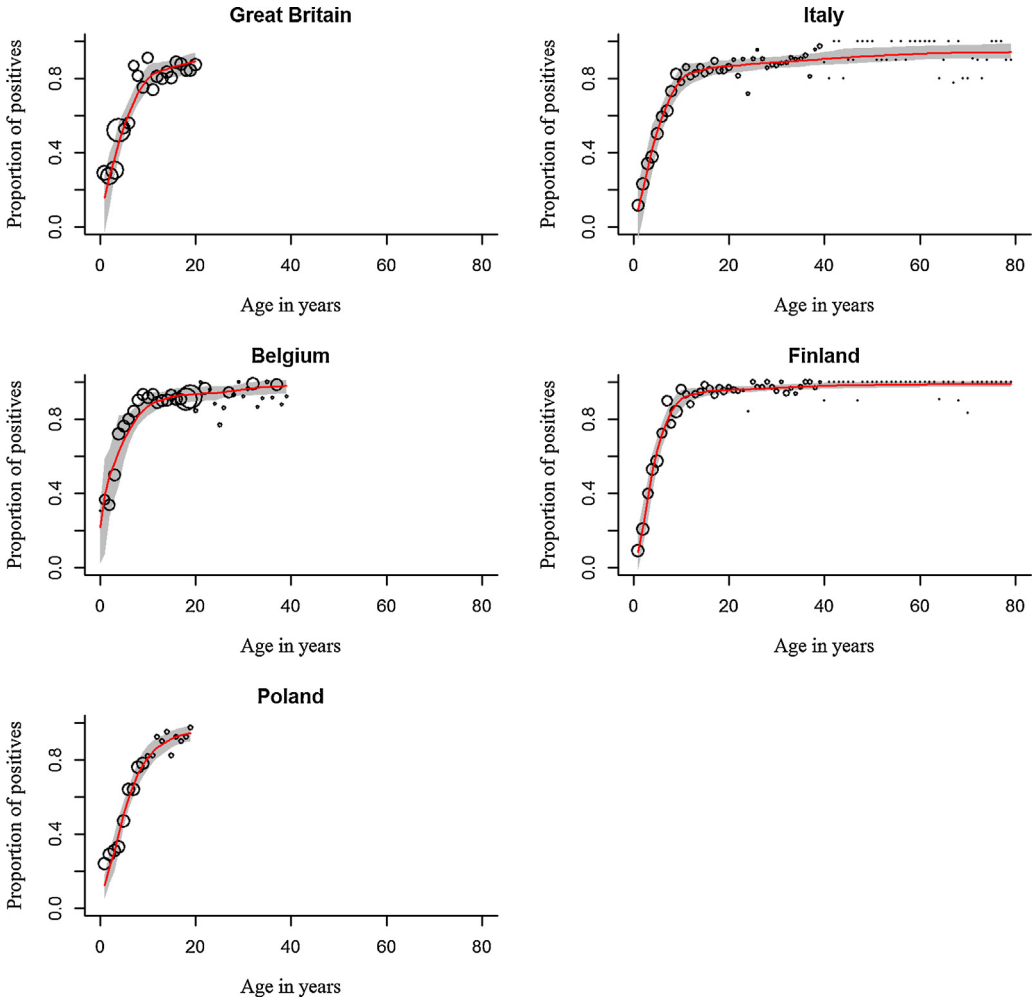


**Fig. 4.** Proportions of sample positive for VZV antibodies by age (blue circle with size proportional to sample size), and model predictions of prevalence (red dash line) by age. The grey lines give the 95% bootstrap interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the contact matrix is somewhat less assortative. Full details are in Appendix A3 (Supplementary data).

## 4. Discussion

Understanding social contact rates is essential to most realistic models of infectious disease transmission. Contact studies such as POLYMOD and its successors have advanced the field by allowing such rates to be measured empirically rather than assumed. However, infection transmission events cannot usually be directly observed so assumptions have to be made about the contribution of different types of reported contacts used as proxies to contacts leading to transmission. We have conducted the first study to investigate the contacts that best explain past infection status data, taking into account more than one characteristic at a time using clustering algorithms.

When these reported contacts are clustered using clustering algorithms, only a few of these clusters are found to be relevant to VZV transmission. VZV transmission in all five countries is dominated by a group of highly assortative contacts and a group of less assortative contacts. However, the balance between the two groups changes across countries: the former group dominates in Finland and Italy, while the latter group dominates in Poland. Great Britain and Belgium lie in between. These country characteristics are robust to constraining the VZV data used to fit models to data from 0 to 20 year olds so they do not appear to be driven by country differences in the availability of VZV seroprevalence data.

Interestingly, although school contacts appear to play a key role across all five countries in providing the assortativeness in contacts, non-assortativeness in contacts derive from home, leisure or other contacts. This matches information from VZV surveillance in England that suggests school and preschool children play a key role in transmission (Brisson et al., 2001). This explanation has face validity since VZV is transmitted between, as well as within, households (Organisation for Economic Cooperation and Development, 2008). Of note, the proportion of children 4 years and under enrolled in education in 2006 (when the POLYMOD contact study was conducted) is highest in Belgium, followed by Italy, United Kingdom, Finland and finally Poland. This exactly matches the order in which assortative contacts dominate the clusters relevant to VZV transmission in the five countries, with the exception of Finland. In Finland, although formal education starts late (at age 7), there is high enrolment in publicly subsidised day-care which is not captured in standardised inter-country statistics.

The quality of fit for models based on clusters stratified by multiple variables was only slightly better than for previous models based on contacts stratified by a single variable alone. The similarity between fit quality with single and multiple stratifications of contacts suggests that a single characteristic largely succeeds in capturing the dichotomy between relevant and non-relevant contacts. This is probably because contacts characteristics are highly

associated with each other (e.g. contacts that are physical, long duration, home and frequent at the same time are disproportionately represented among all contacts). Also, by considering several contact characteristics simultaneously, our clustering method gives insights into the infection transmission process that may not be seen when stratifying by a single characteristic. For instance, we can determine the relevance of short duration physical contacts for which single characteristic stratification would produce conflicting conclusions depending on whether the stratification was by duration or proximity.

Recent analyses have shown that using an age-specific proportionality factor enables models to better fit VZV serology in many European countries (Goeyvaerts et al., 2010; Santermans et al., 2015). In our analysis, we vary proportionality factors according to other contact characteristics, but assume that they are age-independent within each cluster of contacts. This was done for simplicity and to allow direct comparison with previous work (Melegaro et al., 2011) but may have resulted in a poorer fit to data.

## 5. Conclusions

Using the $k$ medoids clustering algorithms to identify clusters of contacts relevant to VZV transmission gives similar model fit quality and $R_0$ values as contacts stratified by a single characteristic. However, considering several characteristics simultaneously provides key insights into the type of contacts that are most relevant to infection transmission, and those that seem not to play an important role. Firstly, it confirms that single-characteristic stratifications are able to provide optimal fits to data, probably because they are able to capture the most intimate contacts responsible for most of transmission. Secondly, we find that school-age children play a key role in almost all contacts relevant to VZV transmission across five countries, and contacts at school in particular are most important in countries with highly assortative contact matrices. This kind of analysis may therefore provide an alternative or supplementary approach to previous single characteristic stratification models, as well as validation for previous methods of contact stratification. Extending such analyses to other countries and infections may be useful for future work.

## Conflicts of interest

None.

## Author contributions

MJ and XX designed the study. XX carried out the analysis, wrote the codes and drew the conclusions with input from MJ, AJvH, MGK and AM. XX and MJ wrote the first draft of the manuscript. All authors contributed to critical revisions of the manuscript and approved the final article.

## Acknowledgments

We thank John Edmunds, Mirjam Kretzschmar and Rafaeal Mikolajczyk for helpful discussions. This work was supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Immunisation at the London School of Hygiene and Tropical Medicine in partnership with Public Health England (PHE). The views expressed are those of the author and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England. Research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement 283955 (DECIDE) to AM. The funders did not influence study design, collection, analysis and interpretation of data, writing of this report, or the decision to submit for publication.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.epidem.2016.08.001.

## References

Beutels, P., Shkedy, Z., Aerts, M., Van Damme, P., 2006. Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface. Epidemiol. Infect. 134, 1158–1166, http://dx.doi.org/10.1017/S0950268806006418.

Breckenridge, J., 2000. Validating cluster analysis: consistent replication and symmetry. Multivariate Behav. Res. 35, 261–285.

Brisson, M., Edmunds, W.J., Law, B., Gay, N.J., Walld, R., Brownell, M., Roos, L.L., Roos, L., De Serres, G., 2001. Epidemiology of varicella zoster virus infection in Canada and the United Kingdom. Epidemiol. Infect. 127, 305–314.

Edmunds, W.J., O'Callaghan, C.J., Nokes, D.J., 1997. Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. Proc. Biol. Sci. 264, 949–957, http://dx.doi.org/10.1098/rspb.1997.0131.

Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. Cluster analysis. In: Wiley Series in Probability and Statistics. Wiley, Chichester, UK, http://dx.doi.org/10.1007/BF00154794.

Goeyvaerts, N., Hens, N., Ogunjimi, B., Aerts, M., Shkedy, Z., Damme, P., Van Beutels, P., 2010. Estimating infectious disease parameters from data on social contacts and serological status. J. R. Stat. Soc. Ser. C Appl. Stat. 59, 255–277, http://dx.doi.org/10.1111/j.1467-9876.2009.00693.x.

Gower, J.C., 1971. A General Coefficient of Similarity and Some of Its Properties. Biometrics 27, 857, http://dx.doi.org/10.2307/2528823.

House, T., Baguelin, M., Van Hoek, A.J., White, P.J., Sadique, Z., Eames, K., Read, J.M., Hens, N., Melegaro, A., Edmunds, W.J., Keeling, M.J., 2011. Modelling the impact of local reactive school closures on critical care provision during an influenza pandemic. Proc. Biol. Sci. 278, 2753–2760, http://dx.doi.org/10.1098/rspb.2010.2688.

Iozzi, F., Trusiano, F., Chinazzi, M., Billari, F.C., Zagheni, E., Merler, S., Ajelli, M., Del Fava, E., Manfredi, P., 2010. Little Italy: an agent-Based approach to the estimation of contact patterns- fitting predicted matrices to serological data. PLoS Comput. Biol. 6, e1001021, http://dx.doi.org/10.1371/journal.pcbi.1001021.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics). John Wiley & Sons.

Keeling, M.J., Rohani, P., 2008. Modeling Infectious Diseases in Humans and Animals. Princeton University Press, http://dx.doi.org/10.1038/453034a.

Kretzschmar, M., Mikolajczyk, R.T., 2009. Contact profiles in eight european countries and implications for modelling the spread of airborne infectious diseases. PLoS One 4, e5931, http://dx.doi.org/10.1371/journal.pone.0005931.

Melegaro, A., Jit, M., Gay, N., Zagheni, E., Edmunds, W., 2011. What types of contacts are important for the spread of infections?: using contact survey data to explore European mixing patterns. Epidemics 3, 143–151, http://dx.doi.org/10.1016/j.epidem.2011.04.001.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G.S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., Edmunds, W.J., 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 5, e74.

Nardone, A., de Ory, F., Carton, M., Cohen, D., van Damme, P., Davidkin, I., Rota, M.C., de Melker, H., Mossong, J., Slacikova, M., Tischer, A., Andrews, N., Berbers, G., Gabutti, G., Gay, N., Jones, L., Jokinen, S., Kafatos, G., de Aragón, M.V.M., Schneider, F., Smetana, Z., Vargova, B., Vranckx, R., Miller, E., 2007. The comparative sero-epidemiology of varicella zoster virus in 11 countries in the European region. Vaccine 25, 7866–7872, http://dx.doi.org/10.1016/j.vaccine.2007.07.036.

Ogunjimi, B., Hens, N., Goeyvaerts, N., Aerts, M., Van Damme, P., Beutels, P., 2009. Using empirical social contact data to model person to person infectious disease transmission: an illustration for varicella. Math. Biosci. 218, 80–87, http://dx.doi.org/10.1016/j.mbs.2008.12.009.

Organisation for Economic Cooperation and Development. (2008). Education at a Glance, Statistics/Education at a Glance/2008:OECD Indicators [WWW Document]. URL http://www.oecd-ilibrary.org/education/education-at-a-glance-2008_eag-2008-en (accessed 7.26.16).

Osborne, K., Gay, N., Hesketh, L., Morgan-Capner, P., Miller, E., 2000. Ten years of serological surveillance in England and Wales: methods results, implications and action. Int. J. Epidemiol. 29, 362–368.

R Core Team. (2012). R: A language and environment for statistical computing.

Santermans, E., Goeyvaerts, N., Melegaro, A., Edmunds, W.J., Faes, C., Aerts, M., Beutels, P., Hens, N., 2015. The social contact hypothesis under the assumption of endemic equilibrium: elucidating the transmission potential of VZV in Europe. Epidemics 11, 14–23, http://dx.doi.org/10.1016/j.epidem.2014.12.005.

Schenzle, D., 1984. An age-structured model of pre- and post-vaccination measles transmission. IMA J. Math. Appl. Med. Biol. 1, 169–191.
Thorrington, D., Jit, M., Eames, K., 2015. Targeted vaccination in healthy school children—can primary school vaccination alone control influenza? Vaccine 33, 5415–5424, http://dx.doi.org/10.1016/j.vaccine.2015.08.031.
Vynnycky, E., White, R., 2010. An Introduction to Infectious Disease Modelling. Oxford University Press, Oxford.
Vyse, A.J., Gay, N.J., Hesketh, L.M., Morgan-Capner, P., Miller, E., 2004. Seroprevalence of antibody to varicella zoster virus in England and Wales in children and young adults. Epidemiol. Infect. 132, 1129–1134.
Walesiak, M., Dudek, A., Dudek, M. (2008). clusterSim: Searching for optimal clustering procedure for a data set.
Wallinga, J., Teunis, P., Kretzschmar, M., 2006. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. Am. J. Epidemiol. 164, 936–944, http://dx.doi.org/10.1093/aje/kwj317.